ETHICAL EXPERT SYSTEMS

Michael S. Victoroff, M.D.
Chair, Bioethics Committee
St. Joseph Hospital, Denver, Colorado

17 East 14th Place
Broomfield, Colorado 80020

Abstract

The title is a double entendre. The discussion approaches expert systems from two directions: "What ethical hazards are created by expert systems in medicine?" and "Would it be ethical to design an expert system for solving problems in bioethics?"

Computers present new ethical problems to society, some of which are unprecedented. These can be categorized under several rubrics. The paper describes a rudimentary scheme for understanding ethical issues raised by computers, in general, and medical expert systems, in particular. It focuses on bioethical implications of AI in medicine; explores norms, assumptions and taboos; and highlights certain ethical pitfalls. Principles are elucidated, for building ethically sound systems. Finally, a proposal is discussed, for the design of an expert system for moral problem solving, and the ethical implications of this notion are analyzed.

Ethical Aspects of Computer Technology

It is valuable to have the opportunity to analyze the moral implications of a new technology, prospectively. When vaccination, penicillin, and CPR appeared, few could foresee their broad ethical consequences. Usually, the import of new therapies is not considered, until they have been employed for a while.

There are a few exceptions. In the early 70's, a moratorium was placed on recombinant DNA research, until more was known about its potential hazards. About the same time, a baby was born in Houston with severe immune deficiency, and was placed in a sterile, plastic bubble, as an experiment in immunology. Subsequent candidates for this treatment were denied, because the natural history of the disease and the effectiveness of the therapy were unknown.

However, these examples are distinguished by their rarity. The usual practice is for researchers to pursue their inquiries, oblivious to the moral consequences the results might have. To some extent, this is proper. For it might be wrong to inhibit free scientific inquiry, with prematurely articulated social concerns.

Yet, a balance needs to be struck, between intellectual freedom; the need to advance the work of civilization; and the equally strong interest of a prudent society, wishing to guard against the recklessness of an amoral technology.

Without trying to say how this balance should be achieved, it seems quite obvious that the field of medical informatics offers abundant material for both sides of the ethical scale. Two things make this field attractive to ethicists who want to observe the gestation - or, perhaps, adolescence - of a new technology.

First, it is relatively apprehendable. Although computers are complicated gizmos, their applications in medical practice are fairly straightforward - so far - and widespread enough to allow a diligent student to get a reasonable overview of the entire field.

Second, it is forseeable. Advances in information technology have proceeded rather predictably, for the last decade or so. For the purposes of sweeping generalizations, the major milestones for the next five years of computer progress are approximately laid out.

So, within general bounds, it is possible to chart the course of medical informatics over the near future, and to discuss some of the influences which this technology may have on the doctor/patient relationship, the conduct of medical practice, and the shape of medical science.

The ethical implications of computers in medicine seem to fall into three conceptual categories:

I. Ethical problems pertaining to computers *qua* simple tools for storing, exchanging, and manipulating data.

II. Ethical problems pertaining to computers *qua* complex devices for reasoning with information.

III. Ethical problems pertaining to computers *qua* sentient beings that make judgements based on values.

Most discussion about the ethics of computer applications in medicine has focused on the first category of problems. Here fall the issues of propriety, warranty, liability, realm of application, regulation, use, and abuse. The privacy issue lies here, and the question of entitlement. Many of these issues are quasi-legal, and not really compelling, philosophically. One interesting exception is whether the nature of a medical diagnostic program can be explained sufficiently well to a lay person to permit a valid informed consent to its use. But, for the most part, the ethical issues of the first class parallel those of other kinds of machinery, such as dialysis machines, or even the

0195-4210/85/0000/0644$01.00 © 1985 IEEE

automobile and the telephone.

The second class of problems deals with the way computers deal with information. To call this "reasoning," deliberately ignores the distinction between data management systems, and true expert systems. While not actually thinking for themselves, computers enhance human reasoning in the solution of complex problems, and, to an extent, channel and guide it. Here, the ethics of computers differ from those of simple tools. While tools are merely powerful, reasoning machines are mysterious.

A screwdriver expands the capabilities of its human user, but not in a mysterious way. Mystery doesn't need to imply the supernatural; certainly there are engineers for whom the workings of computers are completely intelligible, down to the semi-conductor level. "Mysterious" means that the process is "transparent" to the user, in the sense in which software engineers use the term. A machine that makes diagnostic inferences is doing a task on a different order than a screwdriver. To the extent that this occurs outside the human range of perception, a new universe of faith, error, sophistry and deception is introduced to human society. What guards the truth, when there is no common sense?

"Built-in" Values

Every decision entails value judgments. What values are currently being built into our decision support systems?

One of the vices to which all professions fall prey, is to allow a narrow value set to become the yardstick for every decision. This happens in medicine, when a doctor becomes so absorbed in treating disease that the personal interests of the patient are neglected. When a programmer designs an elegant weapon for killing civilians, the same moral violation has occurred.

What are the risks of a system that can predict which patients will develop expensive complications in the hospital, and end up costing the institution money? What about a system that identifies people who have little hope of benefiting from therapy? What about a profile of people who should not be permitted to have children? These uses of information are at least as problematic as the broadcasting of early election returns.

It has been said that, "Information doesn't have ethical implications - only the uses of information can be moral or immoral." This has the same hollowness as the claim that missiles don't kill people; only politicians do. Since Nurenberg, our society has recognized that everyone in a chain of command bears a degree of responsibility for the ultimate consequences of their actions, both intended and unintended. It is clear that this principle has a bearing on systems design.

Patients rightly fear the imposition of an implicit value set which is owned but not acknowledged by the physician, and which colors decisions, without being examined. For example, a physician may take for granted that the patient wants life at any cost, and never inquire whether this is true. What values about health and disease are concealed within our diagnostic software? Is alcoholism a disease or a vice? Is a 1% chance of a cure worth considering? What about a 1% risk of death? How do our treatment protocols reflect the personal, cultural and professional biases of their designers? Should these be made evident to patients?

The dehumanization of the patient, and the imposition of external value systems, are serious and widespread problems in the health care industry - with human doctors being the worst offenders. How can these same doctors, when providing expert input to knowledge engineers, design systems that will practice a better standard of medicine than the designers do themselves?

Ironically, expert systems themselves may provide one solution to this problem. Care can be improved by narrowing the gap between what the clinician knows, and what the clinician actually does. The doctor who says "Okay, Honey, now we're going to have to take out that little old gallbladder of yours," might actually teach his medical students to say, "Mrs. Jones, our tests show that your problems may be caused by your gallbladder. I'd like to suggest some options we have for treatment, and to recommend that your best course, in my opinion, would be to have surgery."

An optimistic view would see an ethically sensitive expert system as one which gives more complete explanations, inquires in depth about beliefs and biases, patiently answers all questions, and never forgets to offer alternatives. It does not cut corners when patients are backed up in the waiting room, or avoid taking a sexual history, just because the patient is the doctor's old elementary school principal. Theoretically, the computer can be as much a force for good as for ill.

But the very consistency that makes expert systems attractive, also guarantees that a problem of inhumane or unethical design will repeat itself systematically. In order to justify any optimism that expert systems can adhere to a high standard of ethical practice, close attention will have to be given to this consideration, in systems design.

The Problem of Too Much Credibility

Another problem is the inherent credibility of anything that comes out of a computer.

People are naturally credulous. Every age in history has its sacred cows, trusted because of what they are, rather than what they do. Today, the sacred cow is the computer. How difficult it has become, to argue a point, without statistics to back it up! Aristotle didn't need a computer to argue that heavy objects have a tendency to fall downward. But today, one would need a complete statistical analysis of thousands of heavy objects, before one could assert the existence of gravity, in a professional journal. It is hard to say just how much this irrational trust in "computer facts" adversely influences human decision making.

Subtler, but more disconcerting to some theoretical physicists, is the fact that "brute force," computation has made it possible to fit formulas and data together that may not belong. Inelegant, but powerful data processing techniques may help validate incorrect theories, when a new formula might fit the facts - if not the data - even better. A generation of mathematical modellers is being trained to depend on mega-computation; perhaps at the cost of intellectual creativity.

People are learning that computers have to be asked questions in special ways, if the answers are going to be meaningful. One can't ask, "What sort of day will it be tomorrow?" But one can ask, "What is the probability of precipitation?" Because of the tendency to trust computers rather than human thinkers, and the fact that most computers can only deal with questions that are asked in a mathematical language, decision makers increasingly tend to ask only questions amenable to this treatment. This leads to the disparagement of arguments, and even the neglect of problems, that don't lend themselves readily to modelling.

(Maybe as people get more like machines, the computers will learn to think more like people. Picture the day when the human asks, "What is the probability of precipitation?" and the computer replies, "With your luck, I'd take an umbrella!")

Undoubtedly, computers will discover new syndromes, reveal unforseen associations, and invent new therapies. But clinicians must be on guard against the knowledge that computers will overlook and dismiss, because of their conceptual limits.

Sentience and Personhood

The ultimate category of moral issues in AI deals with the ethical status of machines as sentient beings. This realm of inquiry irritates many AI researchers. There are at least two legitimate reasons for their discomfort.

First, most people who spend their time with computers are not attuned to the philosophical importance of the ontology of personhood. Second, many of the questions about nonhuman intelligence have been raised from an advocatory standpoint, by people with an axe to grind about the sanctity of the human soul, or some other theological precept. Many debates about artificial intelligence - not in the sense of expert systems but of machine sentience - have been hopelessly chauvinistic, or mired in uninformed dogmatism.

Nevertheless, there are significant philosophical problems to be dealt with, as science contemplates the construction of a computer program that simulates human awareness. What are the responsibilities of the creators, to a race of intelligent beings? Is there anything sacrilegious about respecting intelligence in non-humans? What if a robot passes the Turing test? Could it then own property? Would it be murder, to unplug it, against its will?

The Turing Test has a certain elegance, in the laboratory setting. Opinions differ as to what "intellectual" function this test actually measures, whether it be "thinking," or just "conversing." But from an ethical standpoint, it doesn't matter. The moral question is not about intellectual or conversational abilities, but about personhood.

Personhood is the moral master key to a collection of social rights which are jealously guarded from non-persons. The matter is complicated by the fact that there are several different ways of being a person in society. Personhood is defined legally, medically, ethically, and most important of all, socially. Corporations are persons under the law in certain respects, but not in others. Fetuses have many of the rights of persons in a social sense, but not a legal one. Children are undoubtedly persons in every sense, but their rights are significantly abridged by law and custom.

Long before computers have reached the stage when Asimov's Three Laws of Robotics become more than whimsy, there will be machines that mimic anthropomorphic behavior closely enough to invite treatment as persons. Protagoras said, "Man [sic] is the measure of all things." By this standard, once a sufficient segment of the public decides computers are persons, they will be. (Look for this when natural language processing has advanced about two notches ahead of where it is today.)

Computers and Slaves

In the early 19th century, a debate raged over whether slaves brought from Africa were human, in the same sense as "civilized folk." Granted, they could "simulate" human behavior. Everyone knew that slaves grimaced if you beat them, and it was debated whether this meant that they really felt pain, or merely exhibited an adversive reflex. Slaves went to church, prayed, sang songs, and grew upset when their families were separated, in a touching imitation of real people. The fact that they didn't have souls, and therefore could have no true sense of moral right and wrong, only made more charming their aping of human behavior.

The medical profession was deeply divided on this issue. In order to rationalize behaviors that had the look of intelligence, a variety of physiologic explanations was invoked. The disease "drapetomania" was described in the medical literature: This was a disorder of slaves which came in spasms, like seizures, in which the slave exhibited an uncontrollable desire to run away from its master.

None of these human-like behaviors of slaves was sufficient to convince a large segment of the white population that the slaves were persons. In fact, many abolitionists took up the cause of freeing the slaves in the same spirit that animal advocates object to research on experimental animals.

If our society had this much difficulty deciding whether the slaves were persons, how much hope is there for resolving the question for intelligent machines? The very notion of non-human intelligence is anathema to many scientists, in the same way that evolution was, 100 years ago. It is illuminating to compare the learned debate about slaves in the 1830's, with the correspondence in respected journals today, on animal intelligence.

Chimpanzees can communicate with sign language; dolphins have a complex language of sounds and probably a social culture of some sort; dogs feel emotions, including impatience, joy, frustration, and humor. Dogs also have a finely developed sense of time, can formulate plans and carry them out, dream, worry, and interact with their environment through a rich repertory of vocal and body signals. Whether these behaviors represent "true," as opposed to "simulated" intelligence, is a matter of semantics. The real issue is, what moral duties do we have to entities that act in these ways? Is the gradation in moral status between species based on intellectual prowess, cultural complexity, divine ordination? Or is it just that "might makes right?"

The question of rights for non-human intelligences will be faced with serious urgency during

the next century. Conceivably, in some future age, the homocentric universe will be a quaint curiosity of history, like the geocentric one. Ironically, it may be the advent of reasoning machines that forces humankind to confront the ancient question of the place of humans in the order of nature.

Whether or not the slaves were persons or beasts, the plantation owners were careful to keep them away from weapons. It was clear that the interests of slaves and slave owners were mutually adverse, and the ones in power recognized their vulnerability, if they ever lost the upper hand.

What interests could machines have? Do they like nice clean disk drives, complex problems, or prime numbers? Although it is uncertain what the moral perceptions of an intelligent machine would be, if it were programmed with the same database that we humans have, it is plausible that the machine's interests might be divergent from our own. If this proves to be the case; if intelligent machines do develop a value system incongruent with that of their masters, then it might be a very poor idea to entrust them with the guardianship of all the world's weapon systems.

Expert Systems for Ethical Reasoning

Returning to Category II, what would be the ethical constraints on an expert system for reasoning about moral problems?

Bioethics meets many of the criteria that indicate readiness for an expert system:

1. There are experts in the field who possess specialized knowledge by which they can achieve better results than a non-expert.
2. These experts are in limited supply.
3. There are a reasonable number - but not an excessive number - of basic principles and rules, on which most experts agree.
4. The knowledge in the field is valuable.

It is not clear whether or not there is money to be made in the field of ethical consultation. But malpractice expenses comprise about 1% of the total health care costs in the U.S., which represents 10.6% of the GNP - about $355 billion, lately. And this area is the smallest one in which ethicists have an economic impact on the health care system.

Learning to make mathematical models has taught some of us a lot about logic. Although mathematical models need to be viewed with a great deal more suspiciousness than they generally are, there is no question that the precision of thinking required in making a valid model has been a good discipline for humans to learn. In this spirit, a project can be proposed in the simulation of ethical reasoning.

Nothing finds the whites of people's eyes quicker than telling them of plans to construct an artificial ethicist. Is this just "John Henryism," or is it something deeper? Resistance to the idea of a machine addressing ethical problems comes not only from professional ethicists - who, like other white collar workers, see their domain threatened by automation - but also from a large number of people whose idea of an ethical problem demands that it be addressed by the most humanistic of persons and disciplines, and not by those which are often perceived as being the least, (namely technologists).

This intuition about the sensitivity of ethical decisions, has its parallel in medicine. It recognizes the inherent complexity of human interactions, which have not yet been well enough characterized by a mathematical model to allow machine simulation. Such reservations need to be separated from the Luddite fear of automation that simply sees it as unfair competition and the Devil's work.

It would be folly to propose to replace human moral reasoning with some sort of an "ethical engine." But then, in medicine, nobody is proposing to replace doctors. The stethoscope did not replace ears, it merely enhanced the human ability to hear soft sounds. Similarly, in medicine, computers, especially expert systems, may enhance the clinician's ability to remember fine details, correlate complex relationships, make lengthy calculations, and generally make better decisions, with better data. Some of these functions would help in making better ethical decisions, too.

The problem with most complex ethical decisions, (such as whether to unplug a respirator), is of keeping track of a fairly large number of premises, facts, and variables, that need to be balanced against each other on a number of planes simultaneously. As with chess problems, in some ways, people struggle with ethical problems by fragmenting them into logical paths which are small enough to be followed and analyzed seperately. Of course, the difficulty comes when all of these paths are reintegrated, and "supra paths" and "meta paths" are created, raising the level of complexity by orders of magnitude.

This also happens with ethical argument. It's one thing that gives ethics the undeserved reputation of being imprecise and subjective. This reputation comes not so much from the characteristics of ethics, as the characteristics of people who argue ethical questions badly. Actually, ethics is very much like chess, if one can imagine a chess game in which no one gets to see the last few moves. The number of pieces is legion, and the moves are very complex, but they do obey laws that are, from move to move, quite understandable.

So, one could hardly imagine a better test for an ethicist, than to design software for solving an ethical problem. There are a number of general theories of ethics, each of which contains certain large principles which are fundamental, and a set of rules for applying those principles. A good ethical philosopher practices the discipline of making the steps in an argument clear. But this is a hard skill to teach, and even harder to practice in the heat of a crisis. One way an expert system could help in the teaching of ethics, would be to keep track of the argument as it proceeds, and explain what assumptions and rules were used to arrive at the conclusion.

This explanatory utility would seem to be the most useful feature of an ethical expert system. One could ask the system how much of a change in the weight of a given factor would be necessary to change the outcome of an argument. One could ask the system how the outcome might change, if a different set of assumptions were used. For example, decisions about ordinary and extraordinary medical

treatment would be different, depending on whether a system used the Orthodox Jewish set of values, or those laid down by Pope Pius XII. And, decisions about research on fetuses would be different for someone using a Kantian basis for argument and one using a Millsian approach.

If for nothing more than pedagogical purposes, an expert reasoning machine would allow teachers of ethics to critique arguments much more rigorously. Another use might be to store protocols for investigating ethical problems, such as:

- Making the diagnosis of death
- Testing the validity of an informed consent
- Determining when it might be appropriate to forgo lifesaving treatment
- Deciding when it might be justifiable to override a patient's wishes
- Helping a parent clarify values about the treatment of a defective newborn

Many of the justifications for clinical expert systems can be used for ethical ones, as well. Theoretically, they should enhance the decision making process, improve quality assurance, encourage a uniform standard of care, (where appropriate) and allow the auditing of certain procedures.

Like any advisor, an ethical expert system might not always do as well as the best human experts. But a system could be designed to provide meaningful consultation to the non-expert, and improve performance in the clinical setting.

Ethical "Debugging"

Before any expert medical system is employed, it should be subjected to ethical "debugging." Some tips for avoiding ethical pitfalls can be listed:

1. Has the system passed both Human Subjects Committee (IRB) and Institutional Ethics Committee reviews? (Their functions differ.)
2. Does the system weigh treatment options? If so, check for implicit value judgements.
3. Does the system entail risks? These should be completely and clearly disclosed.
4. Can the system be made available fairly, to all who might reasonably benefit from it?
5. If the system is a prototype, have candidates for trial been selected without discrimination or capriciousness?
6. Does the system operate "transparently," to any extent? Assure human oversight.
7. Was the system designed by specialists? Consider a review by family practitioners.
8. Can the system justify every conclusion, to the user? Auditing should be built in.

Conclusion

The history of medicine is built on the bones of dogmatists. But sometimes doctors forget how accustomed they are to being wrong. Patients have some protections when the doctor gets totally off the track, because of the multiple heuristic safeguards that watch over all decisions. The human operating system is unthinkably more complex than any computer's, and provides error trapping and levels of redundancy, far beyond anything possible in a machine. But the human system works with an enormous amount of "transparency," leaving many of its most important principles undefined and implicit.

Some of the basic ethical assumptions on which a human doctor operates, may become stripped away in the translation to a non-human system. An entire genre of science fiction has evolved around the ironies of behavior that can be imagined, when tiny elements of the human instruction set are omitted from robotic programming. Think of doctors who put finances before healing; of treatments that destroy one organ to save another; of doctors oblivious to pain; who measure success by the number of patients who don't die on their service. Remember the first law for all medical students - often the first forgotten upon graduation: Much of what is "known," is wrong. Can expert systems be designed with the humility to function under this constraint? Even human doctors fail to be human sometimes, in spite of everything. It will be catastrophic, if automation enhances only medicine's most inhuman elements, to the detriment of its primary goals.

The time is near, when expert systems will be trusted with the same, casual confidence accorded to laboratory equipment. Large numbers of patients will come in contact with them, in various settings. Standards of care will emerge, wherein consultation with expert systems will be mandatory.

Someday, a patient will be seriously harmed by an expert system. That patient will sue the doctor, the nurse, the programmer, the system designer, the domain expert, the knowledge engineer, the hardware vendor, the OEM, the maintenance contractor, the power company, and the inventors of the transistor. The burden will be on each of these people to show how their contribution met not only technical standards, but ethical ones, as well.

Artificial intelligence, in all its aspects, offers a broad range of new problems to bioethics. Computers and information technology have already affected our lives in ways that will have reverberations for all time to come. The development of fifth generation systems may cause an upheaval in society, unseen since the development of fire, written language, and the printing press.

Not only the rights and interests of people in contact with the health care industry are at stake, not only privacy, social engineering, and entire professional identities are at stake, but the very foundations of human culture, and our identity as a species, may soon be up for discussion.

There are numerous examples of technologies introduced with a cavalier disregard for their ethical implications. Society is paying a heavy price for some benignly intended, but ill-conceived technological revolutions in medicine today. Researchers on the forefront of AI, are urged to bear in mind the lessons of dialysis, organ transplantation, artificial hearts, D.E.S., the Dalcon shield, in-vitro fertilization, and the swine flu vaccine.

Not all of the social, medical or ethical implications of a new technology can be foreseen. But if society is going to have a chance to preview the ethics of artificial intelligence, now is the time. It would be to everyone's benefit, if an ethical consciousness were evident in the planning phase of each new system, before it was introduced.